密集追踪研究中测验信度的估计：多层结构和动态特性的视角[1]

罗晓慧　刘红云

(北京师范大学心理学部, 应用实验心理北京市重点实验室, 心理学国家级实验教学示范中心(北京师范大学)，北京 100875)

摘　要 随着密集追踪研究在心理学等社会科学领域的广泛运用，密集追踪情境中测验信度的估计也受到越来越多研究者的关注。早期沿用横断研究中信度估计思想或基于概化理论的信度估计方法存在诸多局限，并不适用于密集追踪的情境。针对密集追踪数据的多层结构和动态特性这两大特点，可基于多层验证性因子分析、动态因子分析和动态结构方程模型估计密集追踪研究中测验的信度。通过实证数据的演示与比较，讨论三种估计方法的特点和适用情境。未来研究可基于其它密集追踪模型探讨测验信度的估计，也应重视测验信度的检验与报告。

关键词 密集追踪研究，信度，多层结构，动态特性，动态结构方程模型

分类号 B841

---

[1] 收稿日期**:** 2023-08-29

* 国家自然科学基金项目(32071091)。

通信作者**:** 刘红云, E-mail: hyliu@bnu.edu.cn

1 引言

近年来，密集追踪研究(intensive longitudinal study)在心理学、教育学和管理学等社会科学领域中得到了越来越广泛的运用(Mielniczuk, 2023; Hamaker & Wichers, 2017; Zhou et al., 2021)。这类研究通常采用日记法、经验取样法和生态瞬时评估等方法(Bolger et al., 2003; Bolger & Laurenceau, 2013; Shiffman et al., 2008)收集个体在自然情境中多个时间点的数据(如，20 个时间点以上; Collins, 2006)，相比于传统的回顾性调查和实验室研究，具有低回忆偏差和高生态效度等优势(Bolger et al., 2003; Shiffman et al., 2008; Trull & Ebner-Priemer, 2013)。更重要的是，次数较多且频率较高的密集追踪数据能更精细地捕捉到个体的行为和状态随时间的变化，帮助研究者更深入地探索变量的动态变化过程和变量间的相互作用机制(郑舒方 等, 2021; Hamaker & Wichers, 2017; Zhou et al., 2021)。

虽然密集追踪研究能帮助研究者探索和回答更丰富的研究问题，但它也带来了许多研究方法上的挑战(Hamaker & Wichers, 2017)，密集追踪情境中变量的测量和测验的评估就是其中之一(Mielniczuk, 2023)。以往有大量密集追踪研究采用自我报告的方式测量个体在日常情境中的行为和状态。研究者们通常通过从相应变量的特质测验中选取一道或几道题目并进行一定的改编(如，加入“今天”或“从上次填答至今”等时间提示)来测量该变量的动态变化过程。然而，大部分这类研究都没有对研究所用测验的信度等心理测量学属性进行合理、充分的评估(Stone et al., 2023)。Brose 等人(2020)综述了 2005~2017 年 9 月发表在 *Emotion* 杂志的 50 篇情绪相关密集追踪研究的文章并发现，有 29 篇文章报告了测验的信度，其中仅 10 篇文章明确提到信度的估计基于个体内水平的变异。而在 Trull 和 Ebner-Priemer(2020)对 2012~2018 年发表在心理病理学主要期刊的 63 篇密集追踪研究文章的综述中，仅 30%的文章报告了研究所用测验的心理测量学信息(如信度和效度)。此外，Horstmann 和 Ziegler(2020)对 24 篇关于人格状态的密集追踪研究的梳理发现，大部分研究仅通过将人格特质测验中的题目或形容词转换到状态情境来测量人格的动态变化，而并未预先检验这些测验的信效度，且这些研究中最常见的信度估计方法是先计算每个个体每道题目在所有时刻的平均分，再对整个群体计算题目间一致性作为测验信度的估计。然而，这种方法无法体现人格状态分数的可靠性，并不适用于密集追踪的情境(Horstmann & Ziegler, 2020)。考虑到测验信度的评估是数据分析和结果报告的关键步骤，也是衡量研究结果可靠性的重要依据(叶宝娟 等, 2012; Scherer & Teo, 2020)，有必要针对密集追踪研究的数据特点，提出并采用适宜的信度估计方法。

对密集追踪情境中测验信度估计的早期探索主要分为两类。一类研究沿用横断研究中的信度估计思想，先对密集追踪数据进行聚合或拆分将其转化为类似横断数据的模式，再采用横断研究中常用的信度估计指标(如，α系数)估计密集追踪研究中的测验信度。具体来说，这类研究的信度估计方法包括以下三种(Nezlek, 2017)：(1)先对每个个体每道题目的所有时刻的分数进行聚合(如，求平均分)，再利用这些聚合分数计算测验信度；(2)先将数据按不同时刻进行拆分，再对每个时刻所有个体所有题目的数据分别计算测验信度及其算术平均值；(3)先将数据按不同个体进行拆分，再对每个个体所有题目所有时刻的数据分别计算测验信度及其算术平均值。然而，这些方法都存在一定的局限。比如，第一种方法得到的信度体现的是个体差异分数的可靠性，而非个体动态变化分数的可靠性；第二种方法没有考虑到计算不同时刻的测验信度时用到的被试群体并不相同，得到的多个时刻的测验信度不宜合并；第三种方法忽视了同一个体在不同时刻的作答之间的相互依赖性，这与信度计算中的观测独立性前提假设相违背。综上，这些方法都不适用于密集追踪情境中测验信度的估计。

另一类研究基于概化理论(generalizability theory; Cronbach et al., 1963)提出密集追踪研究中测验信度的估计方法。具体来说，这类研究首先通过确定研究中的测量侧面(facet)来考察测量误差的主要来源，然后采用方差分析估计得到归因于各个测量侧面及其交互作用的方差成分，并基于此计算不同含义的信度。比如，Cranford 等人(2006)认为密集追踪研究中观测分数的变异可以归因于个体、时间和题目这三个测量侧面。随后，他们基于对各个侧面的固定或随机效应的不同假设提出了多种信度计算公式。后续的研究者还将这一方法拓展运用于更多测量侧面的密集追踪情境，并提出了相应的信度计算方法(Schönbrodt et.al, 2021)。然而，基于概化理论的信度估计方法也存在不足(Scherer & Teo, 2020)。比如，这类方法需要满足因子载荷在个体间相等、误差方差随时间不变等较强的假设，而实际数据很难满足这些假设，这就可能导致信度的估计并不准确(Lane & Shrout, 2010)。因此，基于概化理论的信度估计方法也不适用于密集追踪的情境，以往研究也建议不要将基于概化理论的信度估计方法应用于追踪研究中(叶宝娟 等, 2012)。

随着对密集追踪研究认识的不断深化，研究者开始更有针对性地基于密集追踪数据的特点，提出更适用于密集追踪情境的信度估计方法。研究者关注的密集追踪数据的特点主要包括其多层结构和动态特性(Hamaker & Wichers, 2017; Lafit et al., 2021)。密集追踪数据的多层结构通常是指密集追踪的多次重复测量(第一水平)嵌套于个体(第二水平)的数据结构；密集追踪数据的动态特性则是指邻近时间点的观测结果并非相互独立，而是存在一定的关联。聚焦于密集追踪数据的上述两大特点，密集追踪情境中测验信度的估计方法也有了新的进展。

为了帮助对密集追踪研究感兴趣的研究者更好地了解这类研究中测验信度的估计方法，本文将从密集追踪数据的两大特点(即多层结构和动态特性)出发，首先分别介绍聚焦于多层结构的信度估计方法(基于多层验证性因子分析)和聚焦于动态特性的信度估计方法(基于动态因子分析)，然后重点介绍整合了密集追踪数据的多层结构和动态特性的信度估计方法(基于动态结构方程模型)。随后，在实证数据中对这三种信度估计方法进行演示与比较。最后，总结讨论上述三种信度估计方法的特点和适用情境，对相关实践应用提供建议。

2 聚焦多层结构的信度估计方法

基于多层验证性因子分析(multilevel confirmatory factor analysis, MCFA; Geldhof et al., 2014)的信度估计方法聚焦于密集追踪数据的多层结构，在个体内和个体间水平分别估计测验信度，现已广泛运用于发展(Eltanamly et al., 2023; Xu & Zheng, 2022)、教育(Hausen et al., 2023; Neubauer et al., 2022)、社会(Di Sarno et al., 2020; Koval et al., 2019)、临床健康(Gerstberger et al., 2023; Van Der Tuin et al., 2023; Wright et al., 2017)、组织管理(Reis et al., 2016; Schmitt et al., 2017)等心理学领域的密集追踪研究中。

基于多层验证性因子分析的信度估计方法对单维和多维测量结构的情况均适用，本文以单维测量结构的情况为例，多维测量结构的情况可参见以往相关研究(Di Sarno et al., 2020; Neubauer et al., 2022; Wright et al., 2017)。当个体内和个体间水平均为单维测量结构时(如图1)，多层验证性因子分析首先将个体 i 的题目 j 在第 t 个测量时间点的观测分数 Y_{jti} ($j = 1, 2, \ldots, q$; $t = 1, 2, \ldots, T$; $i = 1, 2, \ldots, n$)分解为个体间成分(Y_{ji})和个体内成分($Y^{(w)}_{\mathrm{j}ti}$)：

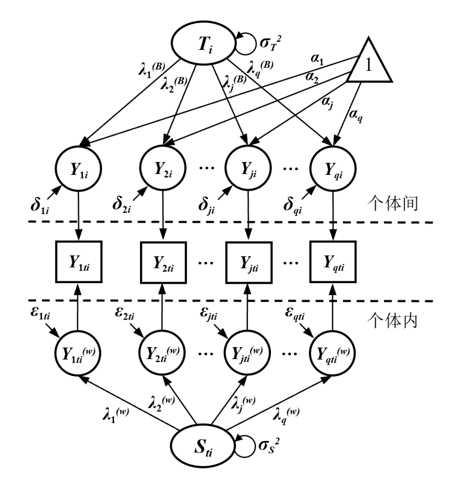


图 1 多层验证性因子分析模型图

注：本图参考了 Geldhof 等人(2014)补充材料中的图 1。

$$Y_{jti} = Y_{ji} + Y_{jti}^{(w)} \tag{1}$$

随后，个体内成分被进一步分解为个体内水平的真分数($\lambda_j^{(w)}S_{ti}$)和误差(ε_{jti})：

$$Y_{jti}^{(w)} = \lambda_j^{(w)}S_{ti} + \varepsilon_{jti} \tag{2}$$

其中，S_{ti}是个体 i 在第 t 个测量时间点的潜在状态因子；$\lambda_j^{(w)}$是题目 j 在个体内水平的因子载荷，对所有个体相等且不随时间变化；ε_{jti}是个体 i 的题目 j 在第 t 个测量时间点的随机测量误差(random measurement error)，假设服从正态分布(即，$\varepsilon_{jti} \sim N(0, \sigma_j^2)$)，各个题目的随机测量误差的协方差为零(即，$cov(\varepsilon_{jti}, \varepsilon_{j'ti}) = 0,\ j \neq j'$)。

个体间成分则被进一步分解为截距(α_j)、个体间水平的真分数($\lambda_j^{(B)}T_i$)和误差(δ_{ji})：

$$Y_{ji} = \alpha_j + \lambda_j^{(B)}T_i + \delta_{ji} \tag{3}$$

其中，α_j是题目 j 的截距；T_i是个体 i 的潜在特质因子；$\lambda_j^{(B)}$是题目 j 在个体间水平的因子载荷；δ_{ji}是个体 i 的题目 j 的测量误差，假设服从正态分布(即，$\delta_{ji} \sim N(0, \sigma_{\delta_j}^2)$)，各个题目的测量误差的协方差为零(即，$cov(\delta_{ji}, \delta_{j'i}) = 0,\ j \neq j'$)。

基于上述模型，可以计算各个题目和各个维度在个体内和个体间水平的信度。在个体内水平，定义某个题目的个体内信度为该题由潜在状态因子解释的变异与该题状态成分的变异之比，各个维度的个体内信度为该维度内各题由潜在状态因子解释的总变异与该维度内各题状态成分的总变异之比。将潜在状态因子的方差固定为 1，可以得到各个题目($Rel_j^{(w)}$)和各个

维度($Rel^{(w)}$)的个体内信度分别为：

$$Rel_j^{(w)} = \frac{(\lambda_j^{(w)})^2}{(\lambda_j^{(w)})^2 + var(\varepsilon_{jti})} \tag{4}$$

$$Rel^{(w)} = \frac{(\sum_{j=1}^{q} \lambda_j^{(w)})^2}{(\sum_{j=1}^{q} \lambda_j^{(w)})^2 + \sum_{j=1}^{q} var(\varepsilon_{jti})} \tag{5}$$

类似地，在个体间水平，定义某个题目的个体间信度为该题由潜在特质因子解释的变异与该题特质成分的变异之比，各个维度的个体间信度为该维度内各题由潜在特质因子解释的总变异与该维度内各题特质成分的总变异之比。将潜在特质因子的方差固定为 1，可以得到各个题目($Rel_j^{(B)}$)和各个维度($Rel^{(B)}$)的个体间信度分别为：

$$Rel_j^{(B)} = \frac{(\lambda_j^{(B)})^2}{(\lambda_j^{(B)})^2 + var(\delta_{ji})} \tag{6}$$

$$Rel^{(B)} = \frac{(\sum_{j=1}^{q} \lambda_j^{(B)})^2}{(\sum_{j=1}^{q} \lambda_j^{(B)})^2 + \sum_{j=1}^{q} var(\delta_{ji})} \tag{7}$$

虽然基于多层验证性因子分析的信度估计方法是密集追踪研究中常用的信度估计方法，但它也存在一定的局限性。比如，这一方法假设各个题目的因子载荷和残差方差对所有个体都相等，故只能得到对个体内水平信度的一个整体评估。然而，这一假设在实际研究中可能并不成立，密集追踪研究中测验信度很可能存在个体间差异(Hu et al., 2016)。此外，基于多层验证性因子分析的信度估计方法没有考虑密集追踪数据中连续观测点之间的时序关系，即忽视了密集追踪数据的动态特性，这可能会影响密集追踪研究中的信度估计结果的准确性。

3 聚焦动态特性的信度估计方法

基于动态因子分析(dynamic factor analysis, DFA)的信度估计方法是密集追踪研究中另一种重要的信度估计方法。动态因子分析最初由 Molenaar(1985)提出，它在 P 技术因子分析(P-technique factor analysis; Cattell et al., 1947)的基础上进一步融入时间序列分析，可以对不同的个体建立不同的模型以考察个体特定(person-specific)的动态过程。后来，有研究者将这一方法应用于密集追踪研究中的信度估计(Fuller-Tyszkiewicz et al., 2017; Lane & Shrout, 2010)。这一信度估计方法能通过考虑变量的自回归过程，体现密集追踪数据的动态特性；还能基于每个个体的数据建立模型，估计个体特定信度，帮助研究者更好地了解不同个体在某个测验信度上的个体间差异。

基于动态因子分析的信度估计方法对每个个体分别建立动态因子模型并计算个体特定信度。类似上述，基于动态因子分析的信度估计方法对单维和多维测量结构的情况均适用，本文以单维测量结构的情况为例(多维的情况可参见 Fuller-Tyszkiewicz 等人(2017)的研究)。

个体 i 的动态因子模型可以分为测量部分和结构部分(如图 2)。测量部分的表达式为：

$$Y_{jti} = \alpha_{ji} + \lambda_{ji} F_{ti} + \varepsilon_{jti} \tag{8}$$

其中，Y_{jti}是个体 i 的题目 j 在第 t 个测量时间点的观测分数($j = 1, 2, \ldots, q$; $t = 1, 2, \ldots, T$; $i = 1, 2, \ldots, n$)；α_j是个体 i 的题目 j 的截距；F_{ti}是个体 i 在第 t 个测量时间点的潜在因子；λ_{ji}是个体 i 的题目 j 的因子载荷；ε_{jti}是个体 i 的题目 j 在第 t 个测量时间点的测量误差，假设服从正态分布(即，$\varepsilon_{jti} \sim N(0, \sigma_{ji}^2)$)，各个题目的测量误差的协方差为零(即，$cov(\varepsilon_{jti}, \varepsilon_{j'ti}) = 0,\ j \neq j'$)。

在结构部分，假设潜在因子满足一阶自回归过程，则结构部分可表示为：

$$F_{ti} = \varphi_i F_{t-1,\, i} + \xi_{ti} \tag{9}$$

其中，φ_i 是个体特定的自回归效应(autoregressive effect)，也被称为惯性(inertia)或滞留效应(carry-over effect)，描述了前一个时间点的潜在因子水平对当前时间点的潜在因子水平的影响；ξ_{ti}是个体 i 的潜在因子在第 t 个测量时间点的动态误差(dynamic error)，假设服从正态分布(即，$\xi_{ti} \sim N(0, \sigma_i^2)$)。

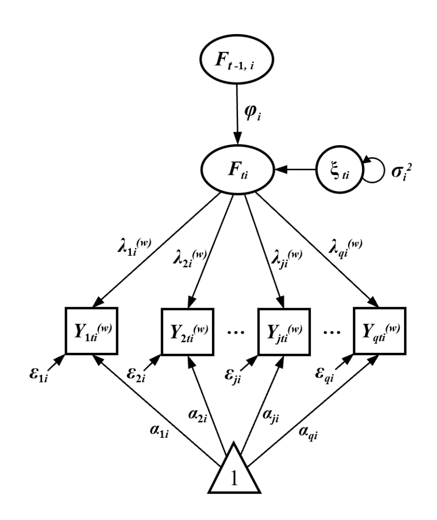


图 2 个体 i 的动态因子模型

基于上述模型，可以计算每个个体的各个题目和各个维度的信度。定义某个题目的个体特定信度为该题由潜在因子解释的变异与该题的总变异之比，各个维度的个体特定信度为该维度内各题由潜在因子解释的总变异与该维度内各题总变异之比。个体 i 的题目 j 的个体特定信度(Rel_{ji})和各个维度的个体特定信度(Rel_i)分别为：

$$Rel_{ji} = \frac{var(\lambda_{ji} F_{ti})}{var(\lambda_{ji} F_{ti}) + var(\varepsilon_{jti})} \tag{10}$$

$$Rel_i = \frac{(\sum_{j=1}^{q} \lambda_{ji})^2 var(F_{ti})}{(\sum_{j=1}^{q} \lambda_{ji})^2 var(F_{ti}) + \sum_{j=1}^{q} var(\varepsilon_{jti})} \tag{11}$$

其中，$var(\lambda_{ji}F_{ti})$是可以由潜在因子解释的变异，等于潜在因子的方差($var(F_{ti})$)与因子载荷的平方(λ_{ji}^2)的乘积；$var(\varepsilon_{jti})$是不可以由潜在因子解释的变异，即测量误差的变异(即，σ_{ji}^2)。由公式(9)可知，潜在因子的方差($var(F_{ti})$)满足下式：

$$var(F_{ti}) = \varphi_i^2 \ var(F_{t-1,\ i}) + var(\xi_{ti}) \tag{12}$$

基于一阶自回归过程的弱平稳假设(weak stationarity assumption)，潜在因子的方差随时间不变(即，$var(F_{ti}) = var(F_{t-1,\ i})$)，故可将公式(12)改写为公式(13)：

$$var(F_{ti}) = \frac{\sigma_i^2}{1 - \varphi_i^2} \tag{13}$$

虽然基于动态因子分析的信度估计方法能估计个体特定信度，还能体现密集追踪数据的动态特性，但它也有一些不足。首先，动态因子分析混淆了观测分数的特质成分(即个体的某一构念在多次观测中的一般水平)和状态成分(即个体的某一构念的某次观测相对其一般水平的偏离)，这可能会导致个体特定信度的估计结果有偏差。其次，这种方法忽视了个体间水平的测量结构，无法估计个体间水平的测验信度。此外，仅利用单一个体的重复测量信息而不考虑其他个体或整个群体的信息可能会导致某些个体模型难以收敛，进而无法估计某些个体的信度(可参见 Fuller-Tyszkiewicz 等人(2017)的研究结果或本文的实证示例)。

4 整合多层结构和动态特性的信度估计方法

基于多层验证性因子分析和基于动态因子分析的信度估计方法都只关注了密集追踪数据的部分特点，而 Asparouhov 等人(2018)提出的动态结构方程模型(dynamic structural equation modeling, DSEM)则为密集追踪数据的多层结构和动态特性的整合提供可能。动态结构方程模型综合了多层模型、时间序列模型和结构方程模型的优势(McNeish & Hamaker, 2020)。它能在个体内和个体间水平分别建立因子模型，考虑变量在不同水平的测量结构，以体现密集追踪数据的多层结构；它还能在个体内水平构建变量的自回归过程，考虑连续观测点之间的时间依赖性，以体现密集追踪数据的动态特性。此外，动态结构方程模型采用贝叶斯估计法，相比于传统的多层模型(采用极大似然估计)可以更灵活地估计参数的随机效应(如参数的个体间差异; McNeish & Hamaker, 2020; Muthén & Asparouhov, 2012)，可以像动态因子模型一样估计得到个体特定信度，故也有研究者将其视为动态因子模型在多层情况下的拓展(Asparouhov et al., 2018)。总之，动态结构方程模型能同时体现密集追踪数据的多层结构和动态特性，还能考察测验信度的个体差异，有助于研究者更好地估计和理解密集追踪研

究中的信度(Luo et al., under review; Xiao et al., 2023)。

类似上述的两种方法，基于动态结构方程模型的信度估计方法对单维和多维测量结构的情况均适用，本文以单维测量结构的情况为例(多维的情况可参见 Xiao 等人(2023)的研究)。对于单维测量结构的构念，常见的两水平动态结构方程模型(two-level DSEM; 如图 3)首先将观测分数分解为个体间成分(即特质成分)和个体内成分(即状态成分)：

$$Y_{jti} = Y_{ji} + Y_{jti}^{(w)} \tag{14}$$

其中，Y_{jti}是个体 i 的题目 j 在第 t 个测量时间点的观测分数($j = 1, 2, \ldots, q$; $t = 1, 2, \ldots, T$; $i = 1, 2, \ldots, n$)；Y_{ji}是个体 i 的题目 j 在所有测量时间点的潜均值(即个体间成分)，代表变量的特质水平；$Y_{jti}^{(w)}$是个体 i 的题目 j 在第 t 个测量时间点的观测分数与该个体在该题目上潜均值的偏

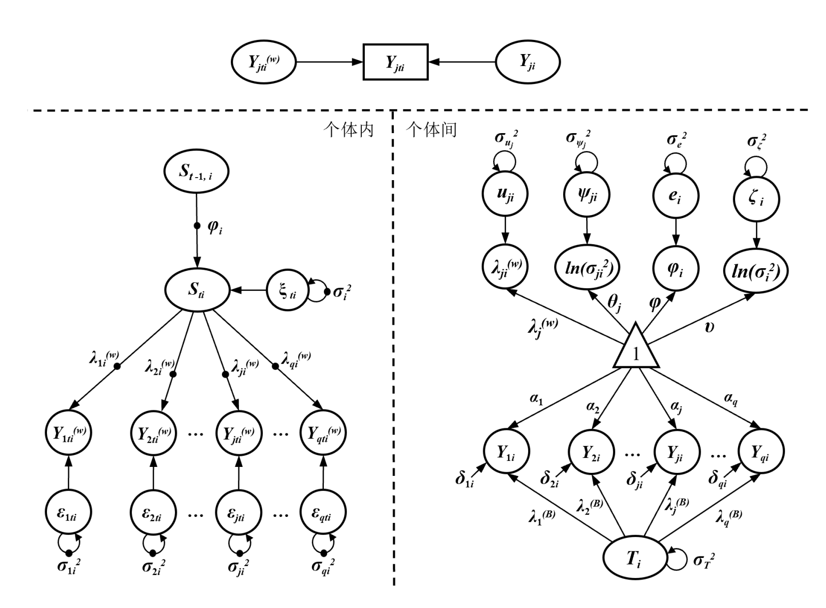


图 3 两水平动态结构方程模型

注：实心圆点代表估计该参数的随机效应，即其个体间差异。本图参考了 Xiao 等人(2023)文章中的图 1。

离值(即个体内成分)，代表变量的状态水平。

然后，对观测分数的个体内成分建立个体内模型(如图 3 左下部分)，包括测量部分和结构部分。在测量部分，个体内成分被进一步分解：

个体内：
$$Y_{jti}^{(w)} = \lambda_{ji}^{(w)} S_{ti} + \varepsilon_{jti} \tag{15}$$

其中，S_{ti}是个体 i 在第 t 个测量时间点的潜在状态因子；$\lambda_{ji}^{(w)}$是个体 i 的题目 j 在个体内水平的因子载荷，在个体间随机估计，假设随时间不变；ε_{jti}是个体 i 的题目 j 在第 t 个测量时间点的随机测量误差，假设服从正态分布(即，$\varepsilon_{jti} \sim N(0, \sigma_{ji}^2)$)，各个题目的随机测量误差的方

差在个体间随机估计，题目间协方差为零(即，$cov\ (\varepsilon_{jti}, \varepsilon_{j'ti}) = 0, j \neq j'$)。

在结构部分，假设潜在状态因子满足一阶自回归过程，公式如下：

个体内： $$S_{ti} = \varphi_i S_{t-1,\ i} + \xi_{ti} \tag{16}$$

其中，φ_i 是个体特定的自回归效应；ξ_{ti}是个体 i 的潜在状态因子在第 t 个测量时间点的动态误差，假设服从正态分布(即，$\xi_{ti} \sim N\ (0, \sigma_i^2)$)。

随后，对观测分数的个体间成分建立个体间模型(如图 3 右下部分)，包括测量部分和随机效应部分。在测量部分，个体间成分被进一步分解：

个体间： $$Y_{ji} = \alpha_j + \lambda_j^{(B)} T_i + \delta_{ji} \tag{17}$$

其中，α_j是题目 j 的截距；T_i是个体 i 的潜在特质因子；$\lambda_j^{(B)}$是题目 j 在个体间水平的因子载荷；δ_{ji}是个体 i 的题目 j 的测量误差，假设服从正态分布(即，$\delta_{ji} \sim N\ (0, \sigma_{\delta_j}^2)$)，各个题目的测量误差的协方差为零(即，$cov\ (\delta_{ji}, \delta_{j'i}) = 0,\ j \neq j'$)。

在随机效应部分，个体内水平的因子载荷($\lambda_{ji}^{(w)}$)、随机测量误差方差的自然对数($ln(\sigma_{ji}^2)$)、自回归效应(φ_i)和动态误差方差的自然对数($ln(\sigma_i^2)$)都被分解为固定部分($\lambda_j^{(w)}$、θ_j、φ和υ)和随机部分(u_{ji}、ψ_{ji}、e_i和ζ_i)：

$$\lambda_{ji}^{(w)} = \lambda_j^{(w)} + u_{ji} \tag{18}$$

$$ln(\sigma_{ji}^2) = \theta_j + \psi_{ji} \tag{19}$$

$$\varphi_i = \varphi + e_i \tag{20}$$

$$ln(\sigma_i^2) = \upsilon + \zeta_i \tag{21}$$

这些个体特定参数的固定部分表示该参数在所有个体间的均值，随机部分表示某一个体对这一均值的偏离值。假设每个参数的随机部分都满足正态分布(即，$u_{ji} \sim N\ (0, \sigma_{u_j}^2)$、$\psi_{ji} \sim N\ (0, \sigma_{\psi_j}^2)$、$e_i \sim N\ (0, \sigma_e^2)$和$\zeta_i \sim N\ (0, \sigma_\zeta^2)$)。值得说明的是，对随机测量误差方差和动态误差方差取自然对数，主要是为了确保估计得到的每个个体的随机测量误差方差和动态误差方差均为正值。此外，对这些误差方差取自然对数还有助于基于多元正态分布，考察这些误差方差的随机对数与其它参数(如个体均值或自回归效应)的随机效应的相关关系(Hamaker et al., 2018)。

基于上述模型，可以计算各个题目和各个维度在个体内和个体间水平的信度。在个体内水平，定义某个题目的个体特定信度为该题由潜在状态因子解释的变异与该题状态成分的变

异之比，各个维度的个体特定信度为该维度内各题由潜在状态因子解释的总变异与该维度内各题状态成分的总变异之比。个体 i 的题目 j 的个体特定信度($Rel_{ji}^{(w)}$)和各个维度的个体特定信度($Rel_{i}^{(w)}$)分别为：

$$Rel_{ji}^{(w)} = \frac{var(\lambda_{ji}^{(w)} S_{ti})}{var(\lambda_{ji}^{(w)} S_{ti}) + var(\varepsilon_{jti})} \tag{22}$$

$$Rel_{i}^{(w)} = \frac{(\sum_{j=1}^{q} \lambda_{ji}^{(w)})^2 var(S_{ti})}{(\sum_{j=1}^{q} \lambda_{ji}^{(w)})^2 var(S_{ti}) + \sum_{j=1}^{q} var(\varepsilon_{jti})} \tag{23}$$

其中，$var(\lambda_{ji}^{(w)} S_{ti})$是可以由潜在状态因子解释的变异，等于潜在状态因子的方差($var(S_{ti})$)与个体内水平的因子载荷的平方($(\lambda_{ji}^{(w)})^2$)的乘积；$var(\varepsilon_{jti})$是不可以由潜在状态因子解释的变异，即随机测量误差的变异(即，σ_{ji}^2)。值得注意的是，潜在状态因子的方差的计算公式与动态因子模型中潜在因子的方差相同，即：

$$var(S_{ti}) = \frac{\sigma_i^2}{1 - \varphi_i^2} \tag{24}$$

此外，通过整合所有个体在各个题目和各个维度的个体特定信度可以分别得到各个题目和各个维度的个体内信度，即描述个体内水平信度的整体指标(具体计算方法见本文的实证示例)。

在个体间水平，定义某个题目的个体间信度为该题由潜在特质因子解释的变异与该题特质成分的变异之比，各个维度的个体间信度为该维度内各题由潜在特质因子解释的总变异与该维度内各题特质成分的总变异之比。题目 j 的个体间信度($Rel_{j}^{(B)}$)和各个维度的个体间信度($Rel^{(B)}$)分别为：

$$Rel_{j}^{(B)} = \frac{var(\lambda_{j}^{(B)} \mathrm{T}_i)}{var(\lambda_{j}^{(B)} \mathrm{T}_i) + var(\delta_j)} \tag{25}$$

$$Rel^{(B)} = \frac{(\sum_{j=1}^{q} \lambda_{j}^{(B)})^2 var(T_i)}{(\sum_{j=1}^{q} \lambda_{j}^{(B)})^2 var(T_i) + \sum_{j=1}^{q} var(\delta_j)} \tag{26}$$

其中，$var(\lambda_{j}^{(B)} \mathrm{T}_i)$是可以由潜在特质因子解释的变异，等于潜在特质因子的方差($var(T_i)$)与个体间水平的因子载荷的平方($(\lambda_{j}^{(B)})^2$)的乘积；$var(\delta_j)$是不可以由潜在特质因子解释的变异，即测量误差的变异(即，$\sigma_{\delta_j}^2$)。

5 实证应用

5.1 实证数据与分析方法

本节将在实证数据中演示如何基于多层验证性因子分析、动态因子分析和动态结构方程

模型估计密集追踪研究中各个题目和维度的信度(以单维测验为例，维度信度即为测验信度，M*plus* 语句和 R 代码见 https://osf.io/n2gw7/?view_only=44938b711ff3425a8e65a87cf523a49c)。实证数据为 252 名女大学生连续 34 天报告的日常拖延数据。参考以往研究对日常拖延的测量(Kühnel et al., 2016; Kühnel et al., 2022; Maier et al., 2021; Van Eerde & Venus, 2018)，本研究在 Tuckman(1991)的拖延量表中加入“今天”的时间提示(如，“今天，我不必要地拖延完成工作，即使是重要的工作”)来测量个体的每日拖延情况。本测验共包括 6 道题，被试需要在每晚睡前从 1(“完全不同意”)到 7(“完全同意”)对每道题进行评分。最终，被试的平均填答率为 94.89%。

基于多层验证性因子分析的信度估计可在 M*plus* 中完成。采用稳健极大似然估计(M*plus* 对两水平模型的默认估计方法)得到多层验证性因子分析模型的参数估计值。同时，根据公式(4) ~ (7)，运用 M*plus* 中的 MODEL CONSTRAINT 语句，直接得到个体内和个体间水平各个题目和整个测验的信度估计值和标准误。

基于动态因子分析的信度估计需要在 R 中调用 M*plus* 完成。具体来说，运用 R 中的 MplusAutomation 包(Hallquist & Wiley, 2018)调用 M*plus*，将每个个体的日常拖延数据分别拟合动态因子模型。采用贝叶斯估计法(固定迭代次数为 10000 次，根据 Hamaker 等人(2018)的建议，通过 PSR 和各参数的轨迹图(trace plot)判断模型此时已收敛，下同)得到各个个体的动态因子模型的参数估计值，并运用 SAVEDATA 语句保存计算个体特定信度所需的参数后验分布(由 200 个可信值(plausible values)组成)。随后，根据公式(10)和(11)，在 R 中计算得到每个个体各个题目和整个测验的个体特定信度的后验分布(由 200 个可信值组成)，后验分布的中位数为该个体的某个题目或整个测验的个体特定信度的点估计，基于所有个体的个体特定信度的点估计可以得到该题目或测验的个体特定信度的分布。

基于动态结构方程模型的信度估计需要同时运用 M*plus* 和 R 完成。在 M*plus* 中，采用贝叶斯估计法(固定迭代次数为 10000 次)得到动态结构方程模型的参数估计值。同时，根据公式(25)和(26)，运用 MODEL CONSTRAINT 语句直接得到个体间水平各个题目和整个测验的信度估计值和 95%贝叶斯可信区间的上、下限。为了估计个体特定信度，首先在 M*plus* 中运用 SAVEDATA 语句保存计算个体特定信度所需的参数后验分布(由 200 个可信值组成)。随后，根据公式(22)和(23)，在 R 中计算得到每个个体各个题目和整个测验的个体特定信度的后验分布(由 200 个可信值组成)。类似基于动态因子分析的信度估计法，可以得到每个个体的某个题目或整个测验的个体特定信度的点估计，以及该题目或测验的个体特定信度的分布。

基于动态因子分析或动态结构方程模型估计信度时，除了估计每个个体的个体特定信度，还可以估计得到个体内信度。个体内信度可以作为个体内水平的信度的整体描述，可与基于多层验证性因子分析得到的个体内信度进行比较。为了得到各个题目或整个测验的个体内信度，用 SAVEDATA 语句保存计算个体特定信度所需的参数后验分布(由 200 个可信值组成)后，先计算每次迭代中每个个体各个题目和整个测验的个体特定信度(每个个体各个题目和整个测验分别可计算得到 200 个个体特定信度)，然后对所有个体求平均，得到该题目或测验的个体内信度的后验分布(由 200 个个体内信度组成)，后验分布的中位数为个体内信度的点估计，2.5%和 97.5%分位数分别为个体内信度的 95%贝叶斯可信区间的上、下限。

此外，值得说明的是，在基于动态因子分析和基于动态结构方程模型计算信度时，某些个体的某些迭代结果中潜在(状态)因子方差的估计值可能为负。为了排除这些有问题的迭代结果对信度估计的影响，我们参考 Xiao 等人(2023)的做法，将相应迭代中的个体特定信度替换为缺失值，即不纳入最终对信度的计算。

5.2 结果与讨论

三种方法估计的各个题目和整个测验的个体间信度和个体内信度如表 1 所示。对于整个测验的信度，基于多层验证性因子分析和基于动态结构方程模型得到的个体间信度的估计值相近，个体内信度的估计值相差相对较大，且都低于基于动态因子分析得到的个体内信度。对于各个题目的信度，三种方法的信度估计结果也存在差异。其中，基于多层验证性因子分析和基于动态结构方程模型得到的各个题目的个体间和个体内信度都相对接近，但基于动态因子分析得到的各个题目的个体内信度都高于基于动态结构方程模型得到的结果。值得注意的是，在基于动态因子分析的信度估计过程中，有 145 人的动态因子模型无法拟合(因为该个体估计的方差协方差矩阵不正定等)，故信度估计结果仅基于模型拟合的 107 人(42.46%)的数据。这可能表明上述对基于动态因子分析与基于其它方法估计得到的信度结果的比较存在问题，因为两者所依据的样本并不相同，研究者应谨慎解读相关结果。更重要的是，这也提醒研究者基于动态因子分析的信度估计方法可能在拟合某些个体模型时存在困难甚至无法成功拟合，相应的个体特定信度无法估计，个体内信度的估计结果也可能存在偏差。

表 1 三种方法的个体间信度和个体内信度

	基于多层验证性因子分析		基于动态因子分析[a]	基于动态结构方程模型	
	个体间信度	个体内信度	个体内信度	个体间信度	个体内信度
题目 1	.954 [.929, .979]	.511 [.047, .550]	.649 [.566, .704]	.973 [.961, .985]	.514 [.500, .528]
题目 2	.731 [.631, .831]	.305 [.266, .344]	.472 [.365, .556]	.851 [.802, .900]	.329 [.311, .343]

题目 3	.905 [.864, .946]	.689 [.654, .724]	.753 [.677, .796]	.930 [.908, .952]	.677 [.658, .687]
题目 4	.903 [.854, .952]	.689 [.644, .734]	.733 [.657, .783]	.948 [.930, .966]	.682 [.667, .694]
题目 5	.946 [.922, .970]	.623 [.586, .660]	.747 [.657, .788]	.966 [.952, .980]	.599 [.585, .609]
题目 6	.963 [.939, .987]	.652 [.615, .689]	.747 [.670, .792]	.990 [.982, .998]	.618 [.603, .629]
测验	.982 [.976, .988]	.892 [.882, .902]	.919 [.890, .937]	.990 [.988, .992]	.847 [.840, .852]

注：中括号内为各参数的 95%(贝叶斯)可信区间的上、下限。[a] 动态因子分析中，145 人的模型无法拟合，个体内信度基于剩余 107 人(42.46%)的模型参数估计结果。

此外，比较各个题目的信度估计结果发现，题目 2(“今天，我推迟做出艰难的决定”)在三种信度估计方法中均呈现出最低的个体间和个体内信度。进一步考察基于动态因子分析和动态结构方程模型得到的个体特定信度的分布(见表 2)发现，在两种可以估计个体特定信度的方法中，题目 2 的个体特定信度组成的分布的中位数和均值都明显低于其它题目，这意味着题目 2 在测量拖延的状态成分时与其它题目的内部一致性较低。结合题目 2 的内容进行分析可以为此提供可能的解释。在 Tuckman(1991)的原量表中，题目 2 用于评估个体推迟做出艰难决定的一般倾向。而本研究在题目 2 中加入了“今天”的时间提示，并用其测量个体每天在多大程度上有推迟做出艰难决定的情况。但值得注意的是，个体并不一定每天都会面临艰难的决定。因此，个体有时可能会对这道题的表述感到困惑或难以作答，故题目 2 和其它题目的一致性也较低。

表 2 基于动态因子分析和动态结构方程模型的个体特定信度的分布

	基于动态因子分析 [a]					基于动态结构方程模型				
	最小值	最大值	中位数	均值	标准差	最小值	最大值	中位数	均值	标准差
题目 1	.027	.978	.655	.648	.172	.041	.994	.534	.515	.204
题目 2	.047	1.000	.474	.471	.262	.034	.984	.290	.329	.204
题目 3	.014	1.000	.782	.753	.178	.032	.999	.724	.676	.236
题目 4	.110	1.000	.795	.737	.205	.030	.999	.745	.683	.238
题目 5	.219	1.000	.760	.746	,148	.055	.999	.641	.599	.238
题目 6	.144	1.000	.770	.745	.157	.056	.999	.638	.619	.214
测验	.651	1.000	.931	.919	.055	.296	.976	.891	.847	.123

注：[a] 动态因子分析中，145 人的模型无法拟合，分布描述基于剩余 107 人(42.46%)的个体特定信度。

6 讨论

6.1 三种方法的比较分析

为了帮助研究者更好地了解并选择合适的信度估计方法，本文对可用于密集追踪情境的三种方法的不同特点和主要局限进行归纳总结(见表 3)。一方面，从数据适配度、可估的信

度和估计方法这三个维度来看，基于动态结构方程模型的信度估计方法整合了基于多层验证性因子分析和基于动态因子分析的优势，能充分体现密集追踪数据的多层结构和动态特性，又能在每个个体、个体内和个体间水平分别估计信度，还能采用贝叶斯估计法更灵活地估计模型参数的随机效应，进而考察个体差异相关的问题。但另一方面，从软件需求和运行耗时这两个维度来看，基于动态结构方程模型的信度估计方法需要用到 Mplus 和其它统计软件(如，R)估计信度，且由于模型相对复杂，程序运行所需时间也较长。相比之下，基于多层验证性因子分析的信度估计方法只需 M*plus* 即可完成，语句简明，结果直接，运行高效，在简便性方面存在优势。此外，表 3 还梳理了三种方法的主要局限。

表 3 三种信度估计方法的比较

	基于多层验证性因子分析	基于动态因子分析	基于动态结构方程模型
数据适配度	体现密集追踪数据的多层结构	体现密集追踪数据的动态特性	体现密集追踪数据的多层结构和动态特性
可估的信度	个体内信度和个体间信度	个体特定信度和个体内信度	个体特定信度、个体内信度和个体间信度
估计方法	极大似然估计	贝叶斯估计	贝叶斯估计
软件需求	只需 M*plus* 即可完成	需在 R 中调用 M*plus*	需要 M*plus* 和其它统计软件(如，R)
运行耗时 [a]	可忽略不计(本例中，小于 1s)	较短(本例中，约 10min)	较长(本例中，约 2h)
主要局限	①对数据有较强的假设 ②无法考察信度的个体差异 ③没有考虑数据的动态特性	①混淆特质和状态成分，信度估计不准 ②忽视多层结构，无法估计个体间信度 ③某些个体模型可能无法拟合	① 操作相对复杂，耗时较长，不够简便

注：[a] CPU 型号为 12th Gen Intel(R) Core(TM) i5-12500H，内存参数为 16GB LPDDR5。

考虑到各种方法的特点和局限，本文对不同方法的适用情境提出建议，并整理提出信度估计方法的选择策略流程图(见图 4)。首先，若研究者不关注测验信度的个体差异或个体特定信度，而是侧重于从整体上了解测验在个体内和个体间水平的信度，或研究者已采用合适的方法(如，交叉分类模型；McNeish et al., 2021)验证测验在不同个体之间的测量满足不变性，则研究者可选用基于多层验证性因子分析的信度估计方法，相对简单地检验并报告测验的个体内和个体间信度。其次，若研究者有理由认为不存在个体特质因子(T_i)对题目作答(Y_{jti})的影响，重点考察不同个体测量模型的差异，关注个体特定信度，或研究的被试量较小(甚至

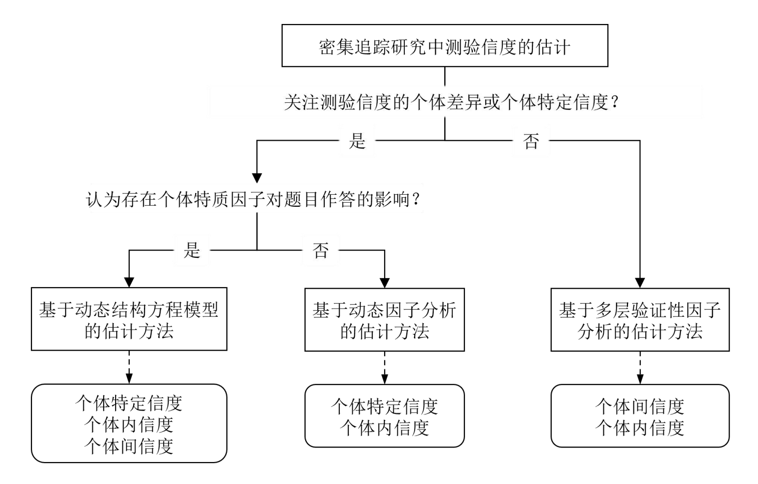


图 4 密集追踪研究中测验信度估计方法的选择策略流程图

是单一个体的时序研究)，不足以考察测验在个体间水平的表现，则研究者可以选择基于动态因子分析的信度估计方法，得到测验的个体特定信度和个体内信度，但此时还需要注意重复测量的时间点是否足够多和个体模型能否成功拟合等问题。然而，在其它大多数情况下，更建议研究者采用基于动态结构方程模型的信度估计方法，得到测验的个体特定信度、个体内信度和个体间信度。现有的许多密集追踪研究通过改编特质测验中的部分题目来测量变量随时间的变化(Horstmann & Ziegler, 2020; Trull & Ebner-Priemer, 2020)，题目的选择和改编效果都缺乏合适的量化分析结果支持。对此，研究者可以基于动态结构方程模型的信度估计方法来充分检验改编后的测验能否可靠地衡量个体间水平的差异以及各个个体和个体内水平整体的动态变化过程。更重要的是，考虑到有研究者呼吁未来研究重视开发更适用于密集追踪情境的测验(Dietrich et al., 2022; Horstmann & Ziegler, 2020; Mielniczuk, 2023)，且近年来也有越来越多这类测验开发的研究(Blanke & Brose, 2017; Engyel et al., 2022; Ringwald et al., 2022)，测验开发阶段的信度估计应该尽可能采用适配于密集追踪数据且可估计各类信度的方法(即，基于动态结构方程模型的信度估计方法)，以帮助测验开发者更好地检验新测验在可靠性方面的表现。

6.2 实践应用中的建议

6.2.1 各个题目的信度

密集追踪研究中各个题目的信度是实践应用中需要关注的一个问题。为了更好地测量研究中的变量，部分研究者会选取多个题目(如，三个及以上)来测量变量随时间的变化过程，但大部分这类研究仅报告整个测验的信度，而未考虑各个题目的信度(Eltanamly et al., 2023; Koval et al., 2019; Van Der Tuin et al., 2023; Wright et al., 2017)。有研究者指出，从特质测验中选取并改编的题目并不一定直接适用于对相应状态的密集测量(Horstmann & Ziegler, 2020; Mielniczuk, 2023)。此外，本文的实证应用也发现，某些改编自特质测验的题目在各种方法得到的各个水平的信度上都低于其它题目，结合题目内容的分析表明，该题可能并不适用于密集追踪的情境。由此可见，应用研究者在检验并报告整个测验的信度之余，还应该进一步考察各个题目的信度。一方面，各个题目的信度估计结果以及题目间的比较分析可以帮助研究者鉴别可能不宜用于密集追踪情境的题目，这对于采用特质测验的改编题测量状态变量的研究尤为重要。另一方面，考虑到有研究者建议在密集追踪研究中采用较短的测验(如，3~6道题；Mielniczuk, 2023)以平衡测验质量和作答负担的影响，对各个题目信度的评估有助于研究者适当缩减密集追踪研究中的测验，提高测量效率。

6.2.2 信度的个体差异

密集追踪情境中值得关注的另一个问题是测验信度的个体差异。在信度的早期研究中有许多研究者强调，信度是一种特定于施测群体的测验特性(Mellenbergh, 1996; Wilkinson, 1999)，基于某个群体得到的信度估计结果不一定能推广到其它群体中。类似地，在密集追踪情境中，研究者关注个体内的动态过程及其测验分数的可靠性，而个体的行为和状态随时间的变化可能具有一定的特异性(Schuurman & Hamaker, 2019)。因此，不同个体的测验分数的可靠性很可能并不相等(Fisher et al., 2018; Stone et al., 2023)，在密集追踪研究中估计测验信度时有必要考虑个体特定信度及其个体间差异。这不仅可以帮助研究者更深入地了解研究所用测验在测量可靠性方面的表现以及对施测群体的适用程度，还可以为研究结果的解读提供更丰富的支持性或警示性信息。对于个体特定信度，大部分个体较高的信度可为个体内水平研究结果的可靠性提供支持，反之，大量个体较低的信度则对个体内水平研究结果的可信度有警示作用，研究者在对相关结果作出解释和推论时需更加谨慎。

6.2.3 信度结果的报告

综合上述两点，我们对密集追踪研究中信度估计结果的报告提出两点建议。首先，考虑到各个题目信度的重要性，建议基于多层验证性因子分析和基于动态结构方程模型估计信度的研究者报告各个题目和整个测验(或各个维度，下同)的个体内信度和个体间信度，基于动态因子分析估计信度的研究者报告各个题目和整个测验的个体内信度，每个信度估计结果应包括其点估计值和(贝叶斯)可信区间的上、下限(参见本文表 1)。这些信度估计结果可以体现研究所用的各个题目和整个测验在个体间和个体内水平的整体表现，有助于识别不适用于密集追踪情境的题目，为测验可靠性评价提供主要参考依据。

此外，对于信度的个体差异问题，如果研究基于动态因子分析或基于动态结构方程模型估计信度，且关注个体特定信度的个体间差异，则研究可以进一步报告个体特定信度的相关结果。具体来说，研究可以呈现各个题目和整个测验的个体特定信度分布图(参见 Xiao 等人(2023)的图 2)，或报告这些分布的描述性统计指标(如中位数、均值和标准差等，参见本文表 2)，以考察题目和测验对各个个体的适用性，为测验可靠性评价提供辅助参考依据。

6.3 其它方法与研究展望

除了已介绍的信度估计方法，在密集追踪情境中测验信度的估计还有需要探索与尝试。比如，受启发于传统的重测信度估计思想，Dejonckheere 等人(2022)通过在密集追踪测验中随机重复一道情绪题，并计算两个分数间的差值平方来估计该题的信度。Hu 等人(2016)还提出可以在密集追踪研究中创建平行测验，并计算每个个体在平行测验上得分的相关来估计

个体特定信度。

此外，还有研究基于潜在特质-状态理论(latent state-trait theory, LST; Steyer et al., 1999, 2015)探讨可用于密集追踪研究的信度估计方法(Castro-Alvarez, Tendeiro, Meijer, & Bringmann, 2022; Castro-Alvarez, Tendeiro, & de Jonge et al., 2022)。潜在特质-状态理论中有三个重要的比例系数(Steyer et al., 2015)：一致性(consistency)、情境特异性(occasion specificity)和可信度(reliability)。一致性是指源于随时间稳定的特质成分的变异与总变异的比例；情境特异性是指源于具体情境的状态成分的变异与总变异的比例；可信度则是一致性和情境特异性之和，即特定情境下源于稳定的特质成分和具体情境的状态成分的变异与总变异的比例，也即随机测量误差之外的变异与总变异的比例。在此理论框架和信度定义下，可以基于多种模型估计密集追踪研究中测验的信度，如多状态-单特质(multistate-singletrait, MSST; Steyer et al., 2015)模型、共同独特特质-状态(common and unique trait–state, CUTS; Hamaker et al., 2017)模型和特质-状态-情境(trait-state-occasion, TSO; Eid et al., 2017)模型。这些模型与本文介绍的模型存在一定关联。比如，多层的共同独特特质-状态模型在统计上与多层验证性因子分析模型相同(Roesch et al., 2010)，而混合效应的特质-状态-情境(mixed-effects trait-state-occasion, ME-TSO; Castro-Alvarez, Tendeiro, & de Jonge et al., 2022)模型与本文介绍的两水平动态结构方程模型在统计上也是相同的。但值得注意的是，由于基于特质状态理论的模型和本文介绍的其它模型在变异分解思路和对信度的定义等方面存在差异，得到的信度估计值及其解释方式也可能不同。对此感兴趣的研究者可以参阅 Castro-Alvarez, Tendeiro, Meijer 和 Bringmann (2022)以及 Castro-Alvarez, Tendeiro 和 de Jonge 等人(2022)的文章。

随着密集追踪研究的不断发展，密集追踪情境中测验信度相关的问题值得更多方法和应用研究者的关注。在方法研究中，测验信度常常基于特定的模型进行估计，而相应信度指标的应用也受限于该模型的适用范围(Laenen et al., 2009)。因此，未来研究可以进一步探讨基于其它模型(如，连续时间结构方程模型; continuous time structural equation modeling, CTSEM; Driver et al., 2017)的信度定义和估计方法。在应用研究中，研究者对测验信度的检验与报告并没有给予足够的重视(Brose et al., 2020; Horstmann & Ziegler, 2020; Stone et al., 2023; Trull & Ebner-Priemer, 2020)，未来研究应该将测验信度的检验作为数据分析的必要步骤，根据具体研究情境选择合适的信度估计方法以得到更合理可靠的研究结论。

参考文献

叶宝娟, 温忠麟, 陈启山. (2012). 追踪研究中测验信度的估计. *心理科学进展, 20*(3), 467–474.

郑舒方, 张沥今, 乔欣宇, 潘俊豪. (2021). 密集追踪数据分析：模型及其应用. *心理科学进展, 29*(11), 1948–1972.

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(3), 359–388.

Blanke, E. S., & Brose, A. (2017). Mindfulness in daily life: A multidimensional approach. *Mindfulness, 8*, 737–750.

Bolger, N., & Laurenceau, J. P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford press.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology, 54*(1), 579–616.

Brose, A., Schmiedek, F., Gerstorf, D., & Voelkle, M. C. (2020). The measurement of within-person affect variation. *Emotion, 20*(4), 677–699.

Castro-Alvarez, S., Tendeiro, J. N., de Jonge, P., Meijer, R. R., & Bringmann, L. F. (2022). Mixed-effects trait-state-occasion model: Studying the psychometric properties and the person–situation interactions of psychological dynamics. *Structural Equation Modeling: A Multidisciplinary Journal, 29*(3), 438–451.

Castro-Alvarez, S., Tendeiro, J. N., Meijer, R. R., & Bringmann, L. F. (2022). Using structural equation modeling to study traits and states in intensive longitudinal data. *Psychological Methods, 27*(1), 17–43.

Cattell, R. B., Cattell, A. K. S., & Rhymer, R. M. (1947). P-technique demonstrated in determining psychophysiological source traits in a normal individual. *Psychometrika*, *12*, 267–288.

Collins, L. M. (2006). Analysis of longitudinal data: The integration of theoretical model, temporal design, and statistical model. *Annual Review of Psychology, 57*, 505–528.

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin*, *32*(7), 917–929.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*(2), 137–163.

Dejonckheere, E., Demeyer, F., Geusens, B., Piot, M., Tuerlinckx, F., Verdonck, S., & Mestdagh, M. (2022). Assessing the reliability of single-item momentary affective measurements in experience sampling. *Psychological Assessment, 34*(12), 1138–1154.

Di Sarno, M., Zimmermann, J., Madeddu, F., Casini, E., & Di Pierro, R. (2020). Shame behind the corner? A daily

diary investigation of pathological narcissism. *Journal of Research in Personality*, *85*, 103924.

Dietrich, J., Schmiedek, F., & Moeller, J. (2022). Academic motivation and emotions are experienced in learning situations, so let's study them [Special issue]. *Learning and Instruction, 81*, 101623.

Driver, C. C., Oud, J. H., & Voelkle, M. C. (2017). Continuous time structural equation modeling with R package *ctsem*. *Journal of Statistical Software, 77*, 1–35.

Eid, M., Holtmann, J., Santangelo, P., & Ebner-Priemer, U. (2017). On the definition of latent-state-trait models with autoregressive effects. *European Journal of Psychological Assessment, 33*(4), 285–295.

Eltanamly, H., Leijten, P., Van Roekel, E., Mouton, B., Pluess, M., & Overbeek, G. (2023). Strengthening parental self‐efficacy and resilience: A within‐subject experimental study with refugee parents of adolescents. *Child Development*, *94*(1), 187–201.

Engyel, M., de Ruiter, N. M., & Urbán, R. (2022). Momentarily narcissistic? Development of a short, state version of the Pathological Narcissism Inventory applicable in momentary assessment. *Frontiers in Psychology, 13*, 992271.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences, 115*(27), E6106-E6115.

Fuller-Tyszkiewicz, M., Hartley-Clark, L., Cummins, R. A., Tomyn, A. J., Weinberg, M. K., & Richardson, B. (2017). Using dynamic factor analysis to provide insights into data reliability in experience sampling studies. *Psychological Assessment*, *29*(9), 1120–1128.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*(1), 72–91.

Gerstberger, L., Blanke, E. S., Keller, J., & Brose, A. (2023). Stress buffering after physical activity engagement: An experience sampling study. *British Journal of Health Psychology, 28*(3), 876–892.

Hallquist, M. N., & Wiley, J. F. (2018). *MplusAutomation*: an R package for facilitating large-scale latent variable analyses in M*plus*. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(4), 621–638.

Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science, 26*(1), 10–15.

Hamaker, E. L., Asparouhov, T., Brose, A., Schmiedek, F., & Muthén, B. (2018). At the frontiers of modeling intensive longitudinal data: Dynamic structural equation models for the affective measurements from the COGITO study. *Multivariate Behavioral Research, 53*(6), 820–841.

Hamaker, E. L., Schuurman, N. K., & Zijlmans, E. A. O. (2017). Using a few snapshots to distinguish mountains

from waves: Weak factorial invariance in the context of trait-state research. *Multivariate Behavioral Research, 52*(1), 47–60.

Hausen, J. E., Möller, J., Greiff, S., & Niepel, C. (2023). Morningness and state academic self-concept in students: Do early birds experience themselves as more competent in daily school life? *Contemporary Educational Psychology, 74*, 102199.

Horstmann, K. T., & Ziegler, M. (2020). Assessing personality states: What to consider when constructing personality state measures. *European Journal of Personality, 34*(6), 1037–1059.

Hu, Y., Nesselroade, J. R., Erbacher, M. K., Boker, S. M., Burt, S. A., Keel, P. K., ... Klump, K. (2016). Test reliability at the individual level. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(4), 532–543.

Koval, P., Holland, E., Zyphur, M. J., Stratemeyer, M., Knight, J. M., Bailen, N. H., ... Haslam, N. (2019). How does it feel to be treated like an object? Direct and indirect effects of exposure to sexual objectification on women's emotions in daily life. *Journal of Personality and Social Psychology*, *116*(6), 885–898.

Kühnel, J., Bledow, R., & Feuerhahn, N. (2016). When do you procrastinate? Sleep quality and social sleep lag jointly predict self - regulatory failure at work. *Journal of Organizational Behavior, 37*(7), 983–1002.

Kühnel, J., Bledow, R., & Kuonath, A. (2022). Overcoming procrastination: Time pressure and positive affect as compensatory routes to action. *Journal of Business and Psychology, 38*(4), 803–819.

Laenen, A., Alonso, A., Molenberghs, G., & Vangeneugden, T. (2009). A family of measures to evaluate scale reliability in a longitudinal setting. *Journal of the Royal Statistical Society Series A: Statistics in Society, 172*(1), 237–253.

Lafit, G., Adolf, J. K., Dejonckheere, E., Myin-Germeys, I., Viechtbauer, W., & Ceulemans, E. (2021). Selection of the number of participants in intensive longitudinal studies: A user-friendly shiny app and tutorial for performing power analysis in multilevel regression models that account for temporal dependencies. *Advances in Methods and Practices in Psychological Science, 4*(1), 1–24.

Lane, S. P., & Shrout, P. E. (2010). Assessing the reliability of within-person change over time: A dynamic factor analysis approach. *Multivariate Behavioral Research*, *45*(6), 1027.

Luo, X., Hu, Y., & Liu, H. (under review). Assessing between- and within-person reliabilities of items and scale for daily procrastination: A multilevel and dynamic approach. *Assessment*.

Maier, T., Kühnel, J., & Zimmermann, B. (2021). How did you sleep tonight? The relevance of sleep quality and sleep-wake rhythm for procrastination at work. *Frontiers in Psychology, 12*, 785154.

McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive

longitudinal data in M*plus*. *Psychological Methods, 25*(5), 610–635.

McNeish, D., Mackinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2021). Measurement in intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(5), 807–822.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1*(3), 293–299.

Mielniczuk, E. (2023). Call for new measures suitable for intensive longitudinal studies: Ideas and suggestions. *New Ideas in Psychology, 68*, 100983.

Molenaar, P. C. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, *50*(2), 181–202.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335.

Neubauer, A. B., Schmidt, A., Schmiedek, F., & Dirk, J. (2022). Dynamic reciprocal relations of achievement goals with daily experiences of academic success and failure: An ambulatory assessment study. *Learning and Instruction*, *81*, 101617.

Nezlek, J. B. (2017). A practical guide to understanding reliability in studies of within-person variability. *Journal of Research in Personality, 69*, 149–155.

Reis, D., Arndt, C., Lischetzke, T., & Hoppe, A. (2016). State work engagement and state affect: Similar yet distinct concepts. *Journal of Vocational Behavior*, *93*, 1–10.

Ringwald, W. R., Manuck, S. B., Marsland, A. L., & Wright, A. G. (2022). Psychometric evaluation of a Big Five personality state scale for intensive longitudinal studies. *Assessment, 29*(6), 1301–1319.

Roesch, S. C., Aldridge, A. A., Stocking, S. N., Villodas, F., Leung, Q., Bartley, C. E., & Black, L. J. (2010). Multilevel factor analysis and structural equation modeling of daily diary coping data: Modeling trait and state variation. *Multivariate Behavioral Research, 45*(5), 767–789.

Scherer, R., & Teo, T. (2020). A tutorial on the meta-analytic structural equation modeling of reliability coefficients. *Psychological Methods, 25*(6), 747–775.

Schmitt, A., Belschak, F. D., & Den Hartog, D. N. (2017). Feeling vital after a good night's sleep: The interplay of energetic resources and self-efficacy for daily proactivity. *Journal of Occupational Health Psychology*, *22*(4), 443–454.

Schönbrodt, F. D., Zygar-Hoffmann, C., Nestler, S., Pusch, S., & Hagemeyer, B. (2021). Measuring motivational relationship processes in experience sampling: A reliability model for moments, days, and persons nested in

couples. *Behavior Research Methods*, *54*(4), 1869–1888.

Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods, 24*(1), 70–91.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology, 4*, 1–32.

Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of states and traits-Revised. *Annual review of clinical psychology, 11*, 71–98.

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality, 13*(5), 389–408.

Stone, A. A., Schneider, S., & Smyth, J. M. (2023). Evaluation of pressing issues in ecological momentary assessment. *Annual Review of Clinical Psychology, 19*, 107–131.

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology, 9*, 151–176.

Trull, T. J., & Ebner-Priemer, U. W. (2020). Ambulatory assessment in psychopathology research: A review of recommended reporting guidelines and current practices. *Journal of Abnormal Psychology, 129*(1), 56–63.

Tuckman, B. W. (1991). The development and concurrent validity of the procrastination scale. *Educational and Psychological Measurement, 51*(2), 473–480.

Van Der Tuin, S., Booij, S. H., Oldehinkel, A. J., Van Den Berg, D., Wigman, J. T. W., Lång, U., & Kelleher, I. (2023). The dynamic relationship between sleep and psychotic experiences across the early stages of the psychosis continuum. *Psychological Medicine*. Advance online publication. https://doi.org/10.1017/S0033291723001459

Van Eerde, W., & Venus, M. (2018). A daily diary study on sleep quality and procrastination at work: The moderating role of trait self-control. *Frontiers in Psychology, 9*, 2029.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604.

Wright, A. G., Stepp, S. D., Scott, L. N., Hallquist, M. N., Beeney, J. E., Lazarus, S. A., & Pilkonis, P. A. (2017). The effect of pathological narcissism on interpersonal and affective processes in social interactions. *Journal of Abnormal Psychology*, *126*(7), 898–910.

Xiao, Y., Wang, P., & Liu, H. (2023). Assessing intra-and inter-individual reliabilities in intensive longitudinal studies: A two-level random dynamic model-based approach. *Psychological Methods*. Advance online

publication. https://doi.org/10.1037/met0000608

Xu, J., & Zheng, Y. (2022). Links between shared and unique perspectives of parental psychological control and adolescent emotional problems: A dyadic daily diary study. *Child Development*, *93*(6), 1649–1662.

Zhou, L., Wang, M., & Zhang, Z. (2021). Intensive longitudinal data analyses with dynamic structural equation modeling. *Organizational Research Methods, 24*(2), 219–250.

Estimating test reliability of intensive longitudinal studies: Perspectives on multilevel structure and dynamic nature

LUO Xiaohui, LIU Hongyun

(Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education (Beijing Normal University), Faculty of Psychology, Beijing Normal University, Beijing, 100875, China)

Abstract

With the widespread use of intensive longitudinal studies in psychology and other social sciences, reliability estimation of tests in intensive longitudinal studies has received increasing attention. Earlier reliability estimation methods drawn from cross-sectional studies or based on generalizability theory have many limitations and are not applicable to intensive longitudinal studies. Considering the two main characteristics of intensive longitudinal data, multilevel structure and dynamic nature, the reliability of tests in intensive longitudinal studies can be estimated based on multilevel confirmatory factor analysis, dynamic factor analysis, and dynamic structural equation models. The main features and applicable contexts of these three reliability estimation methods are demonstrated with empirical data. Future research could explore the reliability estimation methods based on other models, and should also pay more attention to the testing and reporting of test reliability in intensive longitudinal studies.

Key words: intensive longitudinal study, reliability, multilevel structure, dynamic nature, dynamic structural equation modeling